A Knowledge-Based Search Engine Powered by Wikipedia

David Milne Ian H. Witten David M. Nichols

Department of Computer Science, University of Waikato
Private Bag 3105, Hamilton, New Zealand
+64 7 838 4021

{dnk2, ihw, dmn}@cs.waikato.ac.nz

ABSTRACT

This paper describes Koru, a new search interface that offers effective domain-independent knowledge-based information retrieval. Koru exhibits an understanding of the topics of both queries and documents. This allows it to (a) expand queries automatically and (b) help guide the user as they evolve their queries interactively. Its understanding is mined from the vast investment of manual effort and judgment that is Wikipedia. We show how this open, constantly evolving encyclopedia can yield inexpensive knowledge structures that are specifically tailored to expose the topics, terminology and semantics of individual document collections. We conducted a detailed user study with 12 participants and 10 topics from the 2005 TREC HARD track, and found that Koru and its underlying knowledge base offers significant advantages over traditional keyword search. It was capable of lending assistance to almost every query issued to it; making their entry more efficient, improving the relevance of the documents they return, and narrowing the gap between expert and novice seekers.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Search and Retrieval – *search process, query formulation.*

General Terms

Algorithms, Design, Experimentation.

Keywords

Information Retrieval, Query Expansion, Wikipedia, Data Mining, Thesauri.

1. INTRODUCTION

And how will you enquire, Socrates, into that which you do not know? What will you put forth as the subject of your enquiry? And if you find out what you want, how will you ever know that this is the thing that you did not know?

This question, posed by the Greek philosopher Meno some 400 years before Christ's birth, is still relevant in today's internet-savvy age. Whenever we seek out new knowledge—whenever we turn to the ubiquitous search engines—we must grapple with the same fundamental paradox: how can one describe the unknown? That is precisely what must be done to form a query. To make matters worse, search engines are incapable of reasoning with these descriptions as people do. They instead treat a query as nothing more than an excerpt—a few words or phrases—from a relevant document. To search effectively, one must predict not only the information that relevant documents contain, but also the terms by which this is expressed. In short, one must already know a great deal of what is being sought, in order to find it.

What knowledge seekers need—at least those who are not clairvoyant—is a bridge between what they know and what they wish to know, between their vague initial query and the concrete topics and terminology available. One possible bridge is a thesaurus: a map of semantic relations between words and phrases. Knowledge seekers who cannot identify the effective terms for their query could benefit from a thesaurus that covers the terminology of both documents and potential queries, and describes relations that bridge between them. Those who cannot formulate a specific query at all could use a well-organized thesaurus that exposes the topics available and allows them to be explored. The use of thesauri and similar knowledge structures has the potential to greatly advance the art of information retrieval.

In practice, however, thesauri are not widely used to assist with information retrieval. Generic thesauri have shortcomings in any specific technical domain. Domain-specific thesauri are expensive to produce, and their use may require specialist technical knowledge: thus they are only available for a small proportion of document sets, and appeal only to expert users. This research aims to address both issues, by

a) automatically producing thesauri that can serve as a bridge for knowledge seekers, and

b) allowing them to be applied to the searching process intuitively.

This paper focuses on the second goal, the search interface. The next section describes Koru, a search interface that allows a thesaurus, focused to the needs of a particular document collection, to be used intuitively and unobtrusively. This new search system, and its evaluation through a user study, is the main contribution of the paper. However, it cannot work without a comprehensive thesaurus, and Section 3 sketches our new approach to creating thesauri. Section 4 gives some examples of how Koru was used in practice by (untrained) experimental

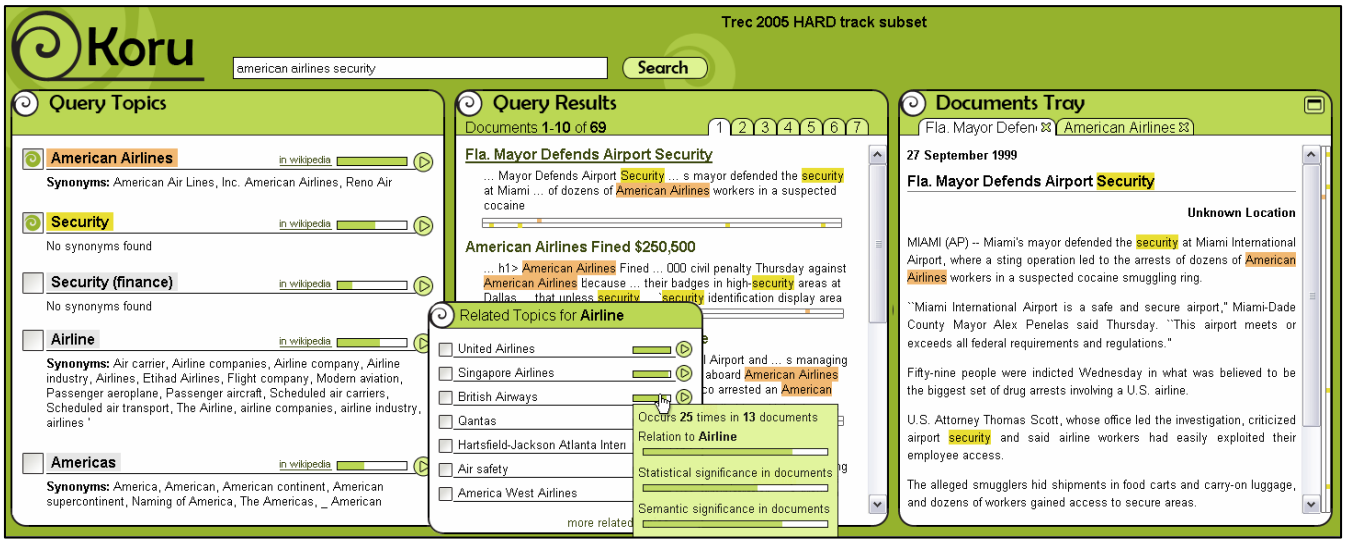

Figure 1: Browsing Koru for topics and documents related to *american airlines security*

subjects. Section 5 describes an evaluation of the system, which to some extent is also an evaluation of the automatically-produced thesaurus. Section 6 presents the context of research surrounding this work, and Section 7 discusses its implications.

2. KORU

Koru is the Māori word for the newborn, unfurling fern frond; a delicate spiral of expanding fractal shapes. For indigenous New Zealanders it symbolizes growth; rebirth; evolution. Likewise, the Koru topic browsing system provides an environment in which users can progressively work towards the information they seek. It exhibits an understanding of the topics involved in both queries and documents, allowing them to be matched more accurately by evolving queries both automatically and interactively.

The interface is illustrated in Figure 1. Implementation is based on the AJAX framework [5], which provides a highly reactive interface couched in nothing more than the standard elements of a webpage. The upper area is a classic search box in which the user has entered the query *american airlines security*. Below are three panels; query topics, query results, and the document tray.

What the figure does not convey is that to avoid clutter not all the panels are visible at any given time. There are three possible configurations, which relate to three stages of expected user behavior:

1. Building an appropriate query. This involves adding and removing phrases until the query and corresponding list of query results satisfies the user's information need. At this stage two panels are visible: query topics and query results (the leftmost two panels in Figure 1).

2. Browsing the document list. Once a suitable list of documents is returned, the user must determine the most relevant ones and judge whether they warrant further study. At this point the panels in Figure 1 slide across so that only the rightmost two—query results and document tray—are visible.

3. In-depth reading of a chosen document. Having located a worthy document, the user then devotes time to actually reading the relevant sections. Here only the documents tray is needed. Anything else would be a distraction.

2.1 The query topics panel

The first panel, query topics, provides users with a summary of their query and a base from which to evolve it. It lists each significant topic extracted from the query, and assigns to each a color that is used consistently throughout the interface. These topics are identified without requiring any special query syntax: in Figure 1 *American Airlines* has been identified as a single phrase even though the user did not surround it by quotes. Sophisticated entity extraction is not used: instead the words and consecutive sequences of words in the query are checked against the thesaurus terms. The only "intelligence" in the process is embodied in the thesaurus and the technique used to generate it.

This thesaurus (described in Section 3) is exceptionally comprehensive. It relates specifically to the document collection and is backed up by a resource that excels in describing contemporary concepts, using contemporary language. Consequently we anticipate that most queries that are valid for the collection will be recognized, even when non-technical terminology or slang is used.[1] However, in the event that terms are not recognized, interaction does not break down: these terms are still listed as topics and incorporated into the query. If the query contains overlapping phrases that each match a thesaurus term, the overlapping words are assigned to the topic with the strongest match against the document collection.

For the given query this results in the five topics *American Airlines*, *Security*, *Security (finance)*, *Airline* and *Americas*. The last is recognized because the thesaurus contains a use-for link from *America* to the preferred term *Americas*. Non-preferred synonyms for each term are listed below that term: For example, the topic *Airline*'s synonyms include *air carrier*, a*irline company,* and *scheduled air transport*. These are used internally to improve queries (see Section 4) and presented to the user in order to help them understand the sense of the topic. The user can also learn more about a topic by clicking the adjacent Wikipedia link.

[1] This expectation is borne out by the experimental evaluation described in Section 5.

Query terms are often ambiguous and relate to multiple entries in the thesaurus. By *security*, for example, the user could also mean property pledged as collateral for a loan, which appears in Figure 1 as *Security (finance)*. Each sense is included, and ranked according to the likelihood that it is a relevant, significant topic for the current query. This likelihood, displayed initially as a horizontal bar next to the topic and elaborated on in a tool-tip, is calculated as a function of a topic's statistical and semantic significance within the document collection. The way in which these weights are obtained is explained in Section 3.1.4.

Only the top-ranked topics that cover all the query terms are used for retrieval (in the example, *American Airlines* and the first meaning of *security*), as indicated by the checkboxes to the left of the topics. This can be overridden manually. For example, it is useful for *Airlines* and *Americas* to appear separately—even though they are not included in Koru's default interpretation of the query—in case the user was interested in all airlines that operated in the U.S. rather than the specific company.

Each topic recognized in the query can be investigated in isolation by using it as a starting point for browsing the thesaurus. In Figure 1 the user has chosen to expand topics related to *Airline*. They have clicked the triangle to the right of that term, which brings up a menu of related topics. They can then investigate further topics of interest such as *Singapore Airlines* and *British Airways*. Any of these topics could be incorporated into the query with a simple click of the appropriate checkbox. As with alternate senses, these topics are ranked according to their expected usefulness, which is elaborated on in tool-tips: the small gray box in Figure 1 shows the tool-tip for *British Airways*. This is calculated in the same way as before, except that the strength of the relation to the parent topic (in this case, *Airline)* is also taken into consideration.

2.2 The query results panel

The second panel in Figure 1, query results, presents the outcome of the query in the form of a series of document surrogates. These resemble those found in typical search engines like Google, and consist of a title and a series of snippets that reflect the document's relationship to the query. Query topics (including synonyms) within both titles and snippets are highlighted for ease of identification.

The only unconventional addition is an overview of how topics are distributed throughout the document, which is presented graphically underneath each snippet using tilebars [13] (only one document in Figure 1 has a fully visible set of tilebars). These represent the entire content of the document as a horizontal bar from left (beginning of document) to right (end). Different bars relate to different query topics, in this case *American Airlines* (upper bar) and *Security* (lower bar). Points, colored in accordance with the query term, appear along the bar to represent the locations in the document of phrases that match the topic. These simple maps can give detailed insights into the relevance of a document. For example, it is apparent that *security* is relevant throughout the first document in Figure 1, but *American Airline*s is mentioned only once. That occurrence is close to a mention of *security*, so the document likely discusses the security of American Airlines, but only in passing. From this purely spatial information the user can make an informed decision about whether the document is worth opening.

2.3 The document tray

The third panel in Figure 1 shows the document tray, which allows the reader to collect multiple documents they wish to peruse. More significantly, its purpose is to facilitate efficient reading by helping users identify relevant sections of a document and navigate between them. These sections are identified using the same information that made the document itself relevant: the query terms used to locate it. Term occurrences are easily seen because they are highlighted according to the colors defined in the query topics panel. Interesting patterns of highlights are likely to indicate sections and paragraphs that should be read.

These highlights can easily be missed, however, because most documents are too large to be viewed without scrolling. Consequently tilebars are supplied to provide an overview of how terms are distributed throughout the document. These tilebars are oriented vertically, and appear on the right-hand side of the standard scrollbar and with a direct mapping to it (they look rather thin in Figure 1). If the scrollbar slider is moved alongside a cluster of points in the tilebar, the highlights that these points represent are visible in the document. Users can jump directly to a particular highlight by clicking the appropriate spot in the tilebar.

3. CREATING A RELEVANT KNOWLEDGE BASE

To work well, Koru relies on a large and comprehensive thesaurus. We took an unconventional approach to obtaining one. Retrieval systems that use thesauri generally use manually-produced ones, either generic (e.g. WordNet [15]) or domain-specific (Agrovoc [7]). Neither are particularly suited to Koru or other information retrieval systems. Generic thesauri are too broad and shallow to provide comprehensive coverage of specific topics, and domain-specific thesauri are expensive to produce and not available in many domains. Another possible route is to use automatically generated thesauri obtained through lexical and statistical analysis of the documents. Unfortunately such natural language processing is quite imprecise and the results tend to be kept behind the scenes. Koru is very transparent in its use of thesauri, and consequently demands a higher level of accuracy because users can easily see its shortcomings.

Manual definition and automatic generation are seemingly exclusive approaches. Our own technique bridges them by automatically extracting thesauri from a huge manually defined information structure. From Wikipedia, we derive a thesaurus that is specific to each particular document collection. Wikipedia is particularly attractive for this work because it represents a vast domain-independent pool of manually defined terms, concepts and relations. By intersecting this with individual document collections, we are able to provide thesauri that are individually tailored to those who seek knowledge from the documents. The intersection operation is necessary because without it an enormous number of links would be presented, most of which would be completely irrelevant to the information retrieval task at hand. The many benefits of such structures, which we call WikiSauri, are covered in [16]. Here we provide an abbreviated sketch of the method by which we derive them.

The basic idea is to use Wikipedia's articles as building blocks for the thesaurus, and its skeleton structure of hyperlinks to determine which blocks are needed and how they should fit together. Each article describes a single concept; its title is a succinct, well-

formed phrase that resembles a term in a conventional thesaurus—and we treat it as such. Concepts are often referred to by multiple terms—e.g. *money* might be grouped with *cash, currency,* and *legal tender*—and Wikipedia handles these using "redirects": pseudo-articles that exist only to connect an alternative title of an article with the preferred one. In earlier work [17] we showed that Wikipedia could provide a viable alternative to Agrovoc [7], a professionally-produced thesaurus for the domain of agriculture, and in particular that Wikipedia redirects match the synonymy encoded in Agrovoc almost perfectly.

The danger in using Wikipedia's structure is that because it is so huge (1 million topics, plus a further 1 million synonyms) the Koru user will become swamped with irrelevant topics and links. It is essential to identify the concepts relevant to a particular document collection, and place these in a structure that allows navigation between related concepts. This requires a measure of semantic relatedness between Wikipedia articles.

3.1.1 Measuring semantic relatedness

Semantic relatedness concerns the strength of the relations between concepts. It can be quantified: for example, one might say that *cash* and *currency* is 100% related, or *currency* and *bank* are 85% related. Despite the evident subjectivity, people are capable of fairly consistent judgments. For example, in [8], 13 participants individually defined relatedness for 350 term pairs and achieved an average correlation of 79% between each individual's judgments and those of the group.

The measure that we use quantifies the strength of the relation between two Wikipedia articles by weighting and comparing the links found within them. Links are weighted by their probability of occurrence; they are less significant for judging the similarity between articles if many other articles also link to the same target. We simply sum the weights of the links that are common to both articles. This yields a correlation of 59% with the above-mentioned manual judgments on the 350 term pairs used in [8].

3.1.2 Disambiguating unrestricted text

To identify the concepts relevant to a particular document collection we work through each document in turn, identifying the significant terms and matching them to individual Wikipedia articles. To lift terms from their surrounding prose, the text is parsed to identify nouns and noun phrases. Candidate concepts for these terms are found in Wikipedia. The fact that it contains redirects and disambiguation pages means this can be done efficiently using only page titles and links.

The sheer scale of Wikipedia makes disambiguation crucial. For example, the term *Jackson* covers over 50 different locations and over 100 different people. If all these were included in the thesaurus, it would become bloated and unfocused. We disambiguate each term using the context surrounding it, using our measure of semantic relatedness to choose the senses that relate most strongly to the other topics in the same sentence. This approach breaks down when the context is insufficient; when there are no unambiguous terms; or if several candidate senses are equally valid. In this case we take a cascading approach: if a sentence contains insufficient information to disambiguate a term the entire surrounding paragraph is used as context; if the paragraph contains insufficient context the entire document is used. It is rare that a term remains ambiguous at the document level, but if so all the equally likely candidate senses are included in the thesaurus.

3.1.3 Identifying relations between concepts

Wikipedia contains many more links than the redirects we use to identify synonymy. It also defines an extensive network of categories that encode hierarchical relations (broader/narrower term, or BT/NT), and millions of hyperlinks between articles which correspond to flat relations (related term RT). These are the links we use to identify related topics, such as the various airlines shown in Figure 1.

Unfortunately the relations in Wikipedia do not map accurately to those in traditional thesauri: categories yield BT/NT relations with only 16% precision, and article hyperlinks are even worse. Consequently we gather all relations from article and category links, but weight them so that only the strongest are emphasized. Moreover, hierarchical and flat relations are not cleanly separated as the structure would suggest, but are intermingled in both category and article links. This is why the Koru interface simply identifies related topics without attempting to specify the nature of the relationship.

3.1.4 Weighting topics, occurrences and relations

Every occurrence of every topic is weighted within the thesaurus. Thus it can be determined whether a document is largely about a topic, or merely mentions it in passing. This is calculated as two weights; standard *tf-idf* (term frequency times inverse document frequency) scores and our own semantic relatedness measure. The former is based on the assumption that a significant topic for a document should occur many times within it, and be useful in distinguishing the document from others. The second is based on the assumption that a significant topic should relate strongly to other topics in the document: here we use the average semantic relatedness measure between a topic and all the others identified for that document.

Some of Koru's functionality depends on these weights. Its ranking of possible senses of query terms (e.g *security* in Figure 1) is based on the significance of each topic within the document collection. This is calculated by aggregating the statistical and semantic significance of all of their occurrences. Koru's ranking of related topics is based on the same measures, plus the strength of the relation between query topic and related topic.

4. KORU IN ACTION

To gain detailed insights into the performance of Koru for document retrieval, we conducted an experiment in which participants performed tasks for which the relevant documents had been manually identified. The tasks, documents and relevance judgments were obtained from the 2005 TREC HARD track [1], which pits retrieval techniques against each other on the task of high-performance retrieval through user interaction. The tasks were specifically engineered to encourage a high degree of interaction.

In order to give a flavor of Koru in action, Table 1 shows three of the TREC tasks, along with information about the initial querying behavior of a few different users for each task. These tasks require the user to think carefully about their query terms, and are unlikely to be satisfied by a single query or document.

The TREC tasks are paired with the AQUAINT text corpus, a collection of newswire stories from the Xinhua News Service, the

New York Times News Service, and the Associated Press Worldstream News Service. The thesaurus that was used throughout was generated using the method described in Section 3; further details are given in Section 5.4.

In the first example in Table 1, User 1 types the query *black bear humans*. Koru the identifies four topics: *American Black Bear*, *Human*, *Bear*, and *Black (people)* (only the first two are shown in Table 1). The first two cover all terms in the query, and are checked by default in the interface. The query that Koru issues to the back-end search engine contains two clauses AND'd together, one for each topic. The first has 4 OR'd components and the second 9, corresponding to synonyms of the topics. Koru places each of these 13 components between quotation marks before passing them to the search engine, so that they are treated as phrases. The result is that a fairly sophisticated query, such as a librarian might issue, has been created from the user's simple three-word input—including some non-obvious synonyms.

User 2 types *black bear man*, which yields precisely the same results. User 3 types *black bear behaviour*, which yields a different query. Notice incidentally how Koru caters for spelling variants and plural forms. Many related topics can be obtained by clicking beside each search topic (as for *Airline* in Figure 1). Examples are *Alaska* and *West Virginia* for the topic *American Black Bear*, *Civilization* for the topic *Human*, and *Psychology*, *Brain* and *Biology* for the topic *Behaviour*.

The second example in Table 1 concerns email abuse. User 1 simply types these two words as the initial query. Each of these terms is recognized as a topic, and behind the scenes Koru automatically expands them to embrace synonyms and alternate forms. User 2 adds the word *employees* which is also recognized as a topic in itself, resulting in a lengthy 3-term query.

In the third example, which is about the Hubble telescope, User 1 types *Hubble telescope achievements*. The first two words are identified as the topic *Hubble Space Telescope*; the word *achievements* is not recognized as a topic at all because it does not appear as a term in the thesaurus. Nevertheless it is still added to the query, along with the expansions of the first topic. User 2 introduces *universe expansion* into the query. Quite fortuitously, the word *expansion* is related in the thesaurus to *Hubble's law* because Wikipedia redirects it to that article: no other senses of *expansion* made it into the thesaurus.

5. EVALUATION

This section describes a user study which evaluated Koru and its underlying data structure for their ability to facilitate and improve information retrieval. Of particular interest is whether the topics, terminology and semantics extracted from Wikipedia make a conclusive, positive difference in the way users locate information, which we measure by pitting the new knowledge-based topic browsing technique against traditional keyword search. We are also interested in Koru's usability; whether it allows users to apply the knowledge found in Wikipedia to their retrieval process easily, effectively and efficiently. This is assessed by observing participants closely as they interact with the system to perform the realistic retrieval tasks provided by TREC.

Example 1: Black Bear Attacks

It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

1 User query *black bears humans*
Main topics *American Black Bear*, *Human*
Query issued
(*American Black Bear* OR *Black Bear* OR *Ursus americanus*)
AND (*Human* OR *All Humankind* OR *Everybody* OR *Homo Sapien* OR *Human Being* OR *Human Kind OR Human species* OR *Humanity* OR *Man*)

2 User query *black bear man*
Same results as above

3 User query *black bear behaviour*
Main topics *American Black Bear, Behavior*
Query issued
(*American Black Bear* OR *Black Bear* OR *Ursus americanus*)
AND (*Behavioral, Behaviors, Behaviour, Behavioural, Behaviours*)

Example 2: Email Abuse

The availability of E-mail to many people through their job or school affiliation has allowed for many efficiencies in communications but also has provided the opportunity for abuses. What steps have been taken by those bearing the cost of E-mail to prevent excesses?

1 User query *email abuse*
Main topics *E-mail, Abuse*
Query issued
(*E Mail* OR *E-Mail* OR *Electronic Mail* OR *E-mail account* OR *Internet mail* OR *Mailto*)
AND (*Abuse* OR *Abused* OR *Abusive* OR *Maltreatment* OR *Mistreatment* OR *Verbal abuse*)

2 User query *email abuse employees*
Main topics *E-mail, Abuse, Employment*
Query issued
the same two clauses as above …
AND (*Employment* OR *Bread and butter* OR *Contract Labour* OR *Employ* OR *Employee* OR *Employer* OR *Job*)

Example 3: Hubble Telescope

Identify positive accomplishments of the Hubble telescope since it was launched in 1991. Documents are relevant that show the Hubble telescope has produced new data, better quality data than previously available, data that has increased human knowledge of the universe, or data that has led to disproving previously existing theories or hypotheses. <further qualifications omitted>.

1 User query *Hubble telescope achievements*
Main topics *Hubble Space Telescope*
Unidentified *achievements*
Query issued
(*Hubble Space Telescope* OR *Hubble Telescope*) AND *achievements*

2 User query *Hubble telescope universe expansion*
Main topics *Hubble Space Telescope, Universe, Hubble's law*
Query issued
(*Hubble Space Telescope* OR *Hubble Telescope*) AND *Universe*
AND (*Hubble's law* OR *Cosmological redshift* OR *Expansion of space* OR *Expansion of the Universe* OR *Hubble Flow* OR *Expansion*)

Table 1: Example retrieval tasks, queries, and topics identified

5.1 Evaluation Procedure

To provide a baseline for comparison we created another version of Koru that provides as much of the same functionality as possible without using a thesaurus, and whose interface is otherwise identical. This allows a clean comparison of the new system with keyword search. The baseline system simply omits the query topics panel in Figure 1. To further reduce interference in the comparison we omitted tilebars from both systems. While they can be of assistance in both topic browsing and keyword searching, they are not a fundamental component of either. We also omitted the Wikipedia links that are placed beside each topic in order to focus the participants on using Koru rather than browsing an external knowledge source.

5.2 Subjects

Twelve participants were observed as they interacted with the two systems. All were experienced knowledge seekers; graduate or undergraduate computer scientists with at least 8 years of computing experience, and all use Google and other search engines daily. Sessions typically lasted for one and a half hours, and were conducted in a controlled environment with video and audio recording, and an observer present. Data was also collected from questionnaires and system logs.

Each user was required to perform 10 tasks (of which Table 1 shows three) by gathering the documents they felt were relevant. Half the users performed five tasks using Koru in one session and the remaining five using the traditional search interface in a second session; for the other half the order was reversed to counter the effects of bias and transfer learning. For each task, approximately 750 relevance judgments are made in which a document is identified as strongly relevant, weakly relevant, or irrelevant.

5.3 Document collection

The ACQUAINT text corpus that was used for the experiments is large—about 3GB uncompressed. It was impractical to create a thesaurus for the entire collection because the process has not been optimized. Instead we used a subset of the corpus: only stories from Associated Press, and only those mentioned in the relevance judgments for the 10 tasks. The result is a collection of approximately 1200 documents concerning a wide range of topics. This was used throughout the experiments.

	Wikipedia	WikiSaurus
Topics	1,110,000	20,000
Terms	2,250,000	57,000
Relations	28,750,000	370,000
Ambiguous document terms		
according to Wikipedia		8500
according to WikiSaurus		3000
Polysemous document topics		
according to documents		2000
according to Wikipedia		6800
according to WikiSaurus		8700

Table 2: Details of Wikipedia and the extracted thesaurus

5.4 Thesaurus

A thesaurus was created automatically for this document collection, based on a snapshot of Wikipedia released on June 3, 2006. The full content and revision history at this point occupy 40 GB of compressed data. We use only the link structure and basic statistics for articles, which consume 500 MB (compressed).

Details of the information available in Wikipedia at this time, and of the thesaurus that was produced, are shown in Table 2. While processing the 1200 documents about 18,000 terms were encountered that matched at least one article in Wikipedia. These are candidates for inclusion in our thesaurus. Including multiple matches yields 20,000 distinct topics—about 2% of those available in Wikipedia.

The disambiguation techniques described in Section 3 greatly reduce the number of multiple matches but do not eliminate them entirely: 47% of terms are ambiguous according to Wikipedia, but this shrank to 17% in the final thesaurus. This residual ambiguity is understandable. Documents in the collection used to derive the thesaurus are not restricted to any particular domain, so terms may well have several valid senses. As an example, the news stories talk of *Apple Corporation's* business dealings and the theft of Piet Mondrian's painting of an *apple* tree.

The full vocabulary of the thesaurus is almost three times larger than the number of topics, because many topics were referred to by multiple terms. 10% of the concepts are expressed by different terms within the document collection itself: e.g. one document talks of *President Bush* and also mentions *George W. Bush*. A further 33% were made so with the addition of Wikipedia redirects: e.g. Wikipedia adds the colloquialisms *Dubya, Shubya* and *Baby Bush* even though these are never mentioned in the (relatively formal) documents. In this context polysemy is desirable, for it increases the chance of query terms being matched to topics and increases the extent to which these are automatically expanded.

The thesaurus was a richly connected structure, with each topic relating to an average of 18 others. As a comparison, Agrovoc [7], a manually-produced and professionally-maintained thesaurus of comparable size, contains just over two relations per topic on average.

5.5 Results

We compared the two systems, Koru and the traditional interface, on the basis of overall task performance, detailed query behavior, and questionnaires that users filled out. In the discussion below we refer to the Koru as "Topic browsing" and the traditional interface as "Keyword searching" because this characterizes the essential difference between the two. Koru identifies topics based on the user's query and encourages topic browsing; the traditional interface provides plain keyword searching.

5.5.1 Task performance

The first question is whether the knowledge base provided by the thesaurus is relevant and accurate enough to make a perceptible difference to the retrieval process. The most direct measure of this is whether users perform their assigned tasks better when given access to the knowledge-based system. Examination of the documents encountered during the retrieval experience shows that this is certainly the case. Table 3 records a significant gain in the recall, precision, and F-measure, averaged over all documents

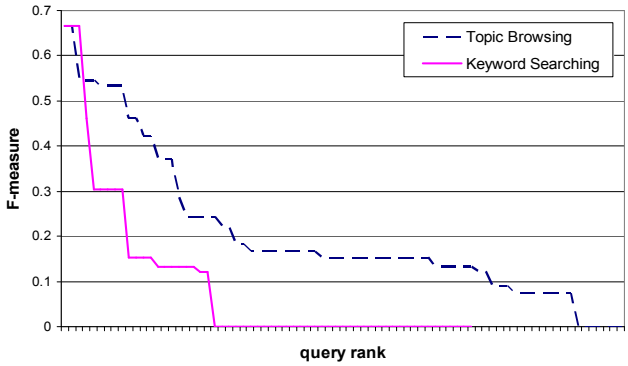

Figure 2: Performance of individual queries

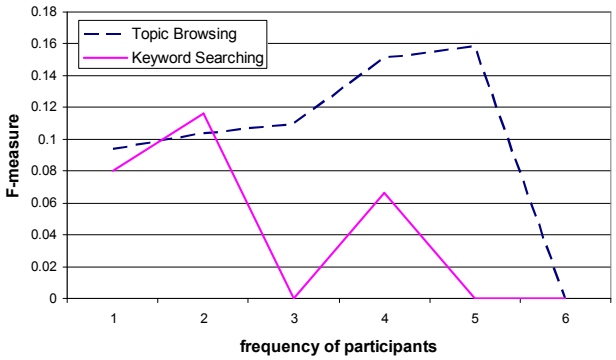

Figure 3: Average performance of queries grouped by number of participants who issued them

encountered using the topic browsing system. This means that the new interface returned better documents than the traditional one.

The greatest gains are made in recall: the proportion of available relevant documents that the system returned. This can be directly attributed to the automatic expansion of queries to include synonyms. Normally gains made in recall are offset by a drop in precision: the inclusion of more terms causes more irrelevant documents to be returned. This was not the case. Table 3 shows no decrease in precision, which attests to the high quality of the Wikipedia redirects from which the additional terms were obtained. Indeed there is even a slight gain, though it is not statistically significant. This can plausibly be attributed to recognition of multi-word terms, which users of traditional interfaces are supposed to encase within quotes. We consistently reminded participants of this syntax when familiarizing themselves with the keyword search interface. Despite this, these expert Googlers did not once use quotation marks, even though they would have been appropriate in 53% of the queries that were issued. The new system performs this often overlooked task reliably and automatically.

Successful topic browsing depends on query terms being matched to entries in the knowledge base. This is typically a bottleneck when using manually defined structures. It is difficult to obtain an appropriate thesaurus to suit an arbitrary document collection, and any particular thesaurus is unlikely to include all topics that might be searched for. Furthermore, specialist thesauri adopt focused, technical vocabularies, and are unlikely to speak the same language as people who are not experts in the domain— the very ones who require most assistance when searching. Koru does not seem to suffer the same problems. For 95% of the queries issued it was able to match all terms in the query (the term *achievements* in Example 3 of Table 1 is a typical exception). We hypothesize that the thesaurus extraction technique provides a knowledge base that is well suited to both the document collection, being grown from the documents, and user queries, being grown from a vocabulary

	Keyword searching	Topic browsing
Recall	43.4%	51.5%
Precision	10.2%	11.6%
F-measure	13.2%	17.3%

Table 3: Performance of tasks

that has been created by both experts and novices. Our user study supports this hypothesis.

5.5.2 Query Behavior

The TREC tasks were specifically selected to encourage user interaction, and participants were invariably forced to issue several queries in order to perform each one. We observed significant differences in query behavior between the two systems.

One major difference was the number of queries issued: 338 on the topic browsing system vs. 274 for keyword searching. This did not correlate to an increase in time spent using Koru, despite its unfamiliarity and greater complexity. Participants were always encouraged to spend 5 minutes on each task regardless of the system used. There are two possible reasons for the increase: Koru either encourages more queries by making their entry more efficient, or requires more queries because they are individually less effective.

Figure 2 indicates that the additional queries are being issued out of convenience rather than necessity. Queries issued by all participants were divided into two groups, one for each interface. Then each group was sorted by F-measure, and the F-measure was plotted against rank. The figure shows that for both topic browsing and keyword searching the best queries had the same F-measure—in other words, the best queries are equally good on both systems. As rank increases a difference soon emerges, however: the performance of keyword searches degrades much more sharply than topic-based ones. In general, the *n*th best query issued when topic browsing is appreciably better (on average) than the *n*th best query issued when keyword searching, for any value of *n*.

This clearly shows that the additional queries issued using Koru are not compensating for any deficiency in performance—for Koru's performance is uniformly better. Instead, it probably reflects the way in which Koru presents the individual topics that make up queries. These are automatically identified and presented to the user, and can be included or excluded from the query with a click of the appropriate checkbox.

We observed several participants modifying their search behavior to take advantage of this feature. They initially issued large, overly specific queries and then systematically selected combinations of the individual terms that were identified. To illustrate this, suppose a user issued a query similar to that in Figure 1 (*american airlines security*) but with additional terms

related to security such as *baggage check, terrorism,* and *x-rays*. This is a poor initial query because few documents will satisfy all topics. But it forms a base for several excellent queries (e.g. *baggage check* and *terrorism,* or *baggage check* and *x-rays*) which in Koru can be issued with a few mouse clicks.

The ability to quickly reformulate queries was greatly appreciated by participants; just under half listed it as one of their favorite features. The only way to emulate this behavior manually in the traditional interface is either by time-consuming re-typing (hence fewer queries issued) or by using Boolean syntax (which even our expert Googlers tended to avoid).

Next we investigate whether it is easier for users to arrive at effective queries when assisted by the knowledge-based approach. In assessing queries we take account of the number of users who made them. A good query issued by many participants is a matter of common sense, whereas one issued by a lone individual is likely to be a product of expert knowledge or some nugget of encountered information.

Figure 3 plots the average F-measure of queries against the number of participants that issued them. At the left are queries issued by only one participant; at the right are ones issued by five and six participants. For the sake of clarity, we have discarded one of the tasks for which the appropriate query terms were particularly easy to obtain. For topic-based queries, performance climbs as they become more common—in other words common queries perform better on average than idiosyncratic ones. This is reversed for keyword searching. Participants were able to arrive at effective queries much more consistently when Koru lent a hand.

The gains we have described are almost exclusively due to automatic query expansion and topic identification. Koru also enables interactive browsing of the topic hierarchy, but we were disappointed to see that participants rarely bothered to use this facility—and even more rarely did such browsing yield additional query topics. In part this was due to users being put off by inaccuracy in the relations that were offered. For example, several participants mentioned that they found it bizarre that Koru identified *homosexuality* as an important topic to investigate if one is interested in *art*. However, this is an exception; typically users felt that the relations were accurate. A more fundamental problem is that even topics that are closely related to a query topic are often irrelevant to the query as a whole. Consider the second example of Table 1, for which most participants issued the query *email abuse.* Most of the related topics for *email* (*browsers, internet, AOL,* etc) and *abuse* (*rape, child abuse, torture,* etc) are perfectly valid but completely irrelevant to the task.

5.5.3 Questionnaire Responses

Each participant completed three separate questionnaires, which solicit their subjective impressions of the two systems. After each session they completed a questionnaire that asked for their impressions of the interface used in that session (Koru or the traditional interface). The third questionnaire was completed at the conclusion of the second session and asked for a direct comparison between the two interfaces, to compare topic browsing and keyword searching directly.

Table 4 shows the results of the final questionnaire, which asked questions like *which of the two systems was more relevant and useful to your needs?* The final question asked participants to name their preferred system overall: two-thirds chose the topic browsing system. Other questions indicate that the main reason for this was relevance and usefulness: in other words the additional functionality that Koru offers is relevant to user needs and produces useful results for their queries. In the words of one participant:

> *The (topic browsing) system provides more choices for users to search for information or documents they need.*

This was somewhat offset by Koru's additional complexity; unsurprisingly, participants felt that the simpler, more familiar system was easier to navigate and use. Simplicity was the reason cited by all participants who chose keyword searching over topic browsing. Several participants took pains to indicate that the difference was marginal. There was no mention of Koru being cumbersome or confusing, just more complex.

> *Not much navigation required (for keyword searching). Topic browsing was very easy to navigate as well.*

> *(Keyword searching is) more minimal. I didn't use the topic browsing stuff anyway.*

The above participant was alluding to Koru's presentation of related topics. As we have already described, this feature was barely used and needs substantial revision. Many participants found it promising however, and two went so far as to list it as their favourite feature.

> *The three different parts (topics, list of articles, one article) are very easy to understand and easy to use. Only the related topics are not so easy to find.*

The remainder of the topic browsing system appeared ergonomic and intuitive for users: there were no other frustrations sited in the surveys and almost all users discovered Koru's full range of features without instruction. We were particularly pleased with the sliding three panel layout. Participants found this unique layout easy to understand and useful, despite its uniqueness and unfamiliarity.

6. RELATED WORK

The central idea of this research is to extract thesauri from Wikipedia and to use them to facilitate query expansion both automatically and interactively. Automatic query expansion is a one step process of adding terms that are synonymous or closely related to those in the query; thus improving recall while hopefully maintaining precision [11]. Interactive query expansion aims to present users with useful terms for exploring new queries and broadening their underlying information needs [18].

Thesaurus-based query expansion is highly dependent on the quality and relevance of the thesaurus. It has been attempted using the manually defined thesaurus structure WordNet [15], both

	Topic	Keyword	Neither
Relevance and usefulness	75.0%	25.0%	0.0%
Ease of navigation	8.3%	66.7%	25.0%
Clarity of structure	41.7%	41.7%	16.7%
Clarity of content	8.3%	41.7%	50.0%
Overall preferred	66.7%	33.3%	0.0%

Table 4: Comparative questionnaire responses

manually [21] and automatically [12], with mixed results. More success has been achieved with automatically generated similarity thesauri [6], which are less accurate but more closely tied to the document collection in question.

However, the best results for query expansion have not been obtained with thesauri at all. The most popular and successful strategy is automatic relevance feedback [3], where terms from the top few documents returned are fed back into the query regardless of any semantic relation. [14] outlines several reasons why individual thesauri can fail to enhance retrieval (e.g. "general-purpose thesauri are not specific enough to offer synonyms for words as used in the corresponding document collection"). Our own research suggests that thesauri extracted from Wikipedia do not suffer the same defects.

A general theme in the literature is the use of external sources to make richer connections between user queries and document collections. Bhogal *et al*. [2] provide a recent review of ontology-based approaches.

Gabrilovich and Markovitch have used external web sources to enhance performance on text categorization tasks [9,10]. Initially they used the hierarchical relationships available from the Open Directory Project[2] and found that their approach was limited by the Project's unbalanced hierarchies and "noisy" text in the web pages that it linked to [9]. In [10] they changed their external source to Wikipedia, with the perceived advantages that its "articles are much cleaner than typical web pages," with larger coverage and more cross-links between articles. Their empirical evaluation "confirmed the value of encyclopedic knowledge for text categorization" and suggested applying similar approaches to other text processing tasks such as information retrieval.

Our work follows this theme but differs in the use of Wikipedia. The essential difference between Gabrilovich and Markovitch's work and our own is that they focus on Wikipedia as a structured collection of documents, while we focus on it as a network of linked concepts and largely ignore the text it contains. Their perspective lends itself to natural language processing techniques, while ours lends itself to graph and thesaurus based ones.

Although a number of interfaces enhanced with thesauri have been developed, few have been evaluated to assess their impact on users' query formulation [20]. In a recent example, Shiri and Revie [19] reported that thesaurus enhancement produced substantially different reactions from university faculty (who commented on narrowing effects) and postgraduate students (who appreciated broadening effects). Their participants also commented on difficulties with AND and OR operators and a dislike of separate term entry. Koru expands queries and permits rapid (re)formulation of queries based on simplified term entry, so we did not encounter responses like this. However, we did experience similar results for topic browsing (Section 5.5.2): where inaccurate additional topic suggestions impeded the ability of participants to develop their queries [19].

It is worth noting that studies such as [19] are often limited by the domain restrictions of the selected thesaurus (in this case agriculture). In principle the Wikipedia-based approach should provide much greater domain coverage and allow future work to use authentic user queries rather than those specifically generated for an evaluation exercise.

[2] http://www.dmoz.org

7. CONCLUSION

This paper has introduced Koru, a new search engine that harnesses Wikipedia to provide domain-independent knowledge-based retrieval. Our intuition that Wikipedia could provide a knowledge base that matched both documents and queries has so far been borne out. We have tested it with a varied domain-independent collection of documents and retrieval tasks, and it was able to recognize and lend assistance to almost all queries issued to it, and significantly improve retrieval performance. Koru's design was also validated, in that it allowed users to apply the knowledge found in Wikipedia to their retrieval process easily, effectively and efficiently. The following quote, given by one participant at the conclusion of their session, summarizes Koru's performance best:

> *It feels like a more powerful searching method, and allows you to search for topics that you may not have thought of…*
>
> *…it could use some improvements but the ability to graphically turn topics on/off is useful, and the way the system compresses synonymous terms together saves the user from having to search for the variations themselves. The ability to see a list of related terms also makes it easier to refine a search, where as with keyword searching you have to think up related terms yourself.*

Koru currently provides automatic query expansion that allows users to express their information needs more easily and consistently. This one-step process of improvement can only take queries so far, however. To go further, one must enter into a dialog with the searcher and interact with them to hone queries and work progressively towards the information they seek. To invoke the imagery offered by Koru's namesake, such a system would allow initial hazy queries to gradually unfold and open out into complete paths across the information space. Our goal in the future is to improve Koru's interactive query expansion facilities until it provides this ability to unfurl queries, thereby living up to its name.

8. REFERENCES

[1] Allan, J. (2005) HARD Track overview in TREC 2005 high accuracy retrieval from documents. *Proc TREC-2005.*

[2] Bhogal, J., Macfarlane, A., and Smith, P. (2007) A review of ontology based query expansion. *Information Processing & Management 43*(4) 866-886.

[3] Billerbeck, B. and Zobel, J. (2004) Questioning query expansion: an examination of behaviour and parameters. *In Proc Australian Database Conf*, pp. 69-76.

[4] Bodner, R. and Song, F. (1996) Knowledge-based approaches to query expansion in information retrieval. *Advances in Artificial Intelligence*, 146-158.

[5] Crane, D., Pascarello E. and James, D. (2005) *Ajax in Action*. Manning, Connecticut.

[6] Curran, J. R. and Moens, M. (2002) Improvements in automatic thesaurus extraction. *Proc. ACL Workshop on Unsupervised Lexical Acquisition*, pp. 59-66.

[7] FAO (1995) *Agrovoc Multilingual Agricultural Thesaurus,* Food and Agricultural Organization of the United Nations.

[8] Finkelstein, L., Gabrilovich, Y.M., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2002) Placing search in context: The concept revisited. *ACM Trans Office Information Systems 20*(1), 406-414.

[9] Gabrilovich, E. and Markovitch, S. (2005) Feature Generation for Text Categorization Using World Knowledge. *Proc. Int Joint Conf on Artificial Intelligence*, pp 1048-1053.

[10] Gabrilovich, E. and Markovitch, S. (2006) Overcoming the Brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. *Proc. American Association for Artificial Intelligence*, pp. 1301-1306.

[11] Greenberg, J. (2001) Automatic query expansion via lexical-semantic relationships, *J American Society for Information Science and Technology 52*(5), 402-415.

[12] Grootjen, F. A. and van der Weide, T. P. (2006) Conceptual query expansion. *Data & Knowledge Engineering 56*(2), 174-193.

[13] Hearst, M.A. (1995) TileBars: visualization of term distribution information in full text information access, *In Proc. SIGCHI Conf on Human Factors in Computing Systems*, pp. 59-66.

[14] Mandala, R., Tokunaga, T., and Tanaka, H. (1999) Combining multiple evidence from different types of thesaurus for query expansion. *In Proc. of SIGIR'99*. 191-197.

[15] Miller, G. A. (1995) WordNet: a lexical database for English. *Communications of the ACM 38*(11), 39-41.

[16] Milne, D. and Witten, I.H. (2007) Extracting Corpus Specific Knowledge Bases from Wikipedia, Working Paper 03/2007, Department of Computer Science, University of Waikato.

[17] Milne, D., Medelyan, O. and Witten, I. H. (2006) Mining Domain-Specific Thesauri from Wikipedia: a case study. *Proc IEEE/WIC/ACM International Conference on Web Intelligence*, Hong Kong, China.

[18] Ruthven, I. (2003) Human interaction: Re-examining the potential effectiveness of interactive query expansion. *Proc. SIGIR Conf on Information Retrieval*, pp. 213-220.

[19] Shiri, A. and Revie, C. (2005) Usability and user perceptions of a thesaurus-enhanced search interface. *Journal of Documentation 61*(5), 640-656.

[20] Shiri, A. and Revie, C. (2006) Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *J American Society for Information Science and Technology 57*(4), 462-478.

[21] Voorhees, E.M. (1994) Query expansion using lexical-semantic relations. *Proc. SIGIR Conf on Information Retrieval*, pp. 61-69.